Death of Disk Panel – HEC FSIO 2011

Nisha Talagala

ntalagala@fusionio.com

Panel Questions

- Is Disk really dying – what about Tape?
- What is the most cost effective use of solid state storage in HEC systems?
- How will disk be placed for HEC IO?
  As capacity devices servicing only well formed IO because the layer above will shape the IO properly?
  If so, why not just use tape?

Non Volatile Memory

FUSION-IO

- Flash
  - 100s GB of NVM per PCIe device
  - SLC for highest performance, MLC for highest capacity
  - Trend of MLC – increase in density, reduction of write cycles
- PCM
  - Still in research
  - Potential of extreme performance increase

750 MB/s
145,000 IOPs
640 GB

ioDrive®

1.5 GB/s
278,000 IOPs
1.28 TB

ioDrive Duo®

6 GB/s
1,180,000 IOPs
5.12 TB

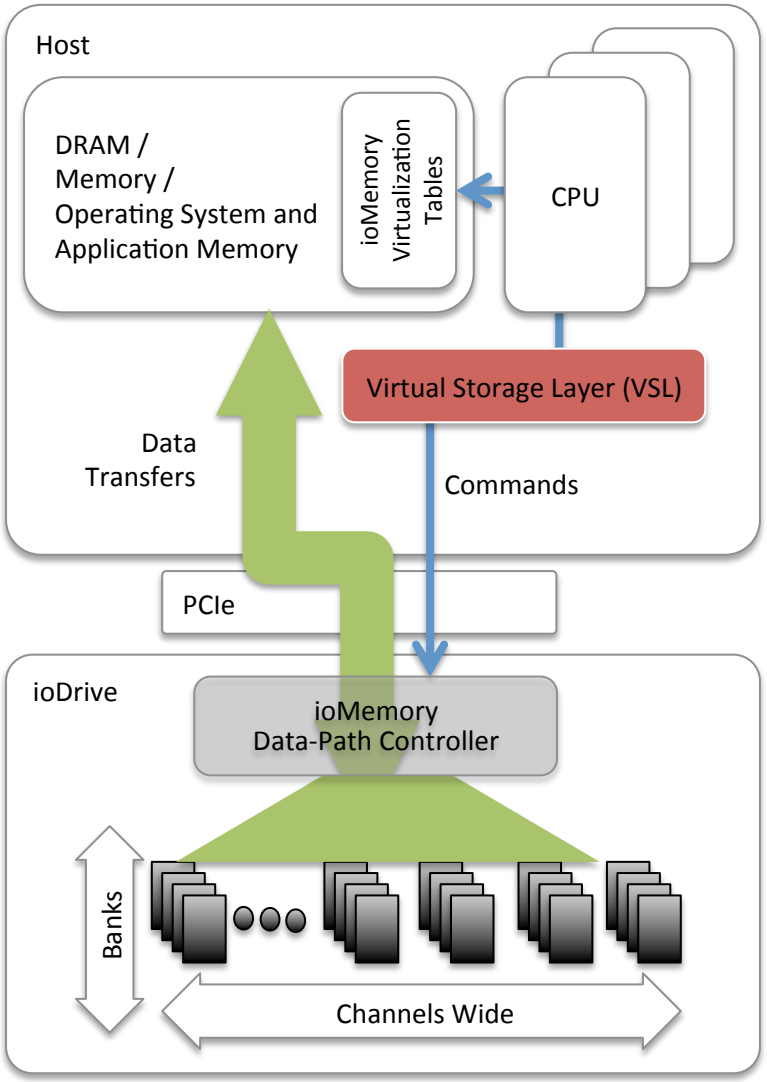
ioDrive® Octal

How to effectively use flash?

- Performance
  - Closer to CPU is best – highest bandwidth, lowest latency
  - Server (compute) side flash complements storage side flash
- Hierarchy of DRAM, flash, disk
- Disk displacement usages
  - Caches – server and storage side
  - Scale out and cluster file systems
    - flash in metadata server
    - storage server
  - Staging, checkpoint
- DRAM displacement usages
  - Improved paging, semi-external memory

Virtual Storage Layer

- Cut-thru architecture – avoids traditional storage protocols
  - Scales with multi-core
- Traditional block access methods for compatibility
- New access methods and primitives natively supported by FTL
  - "DFS", "Beyond Block I/O"
  - Discard (TRIM), Sparse address space, Atomics



[1] http://hpca17.ac.upc.edu/web/?q=node/11

Usage examples

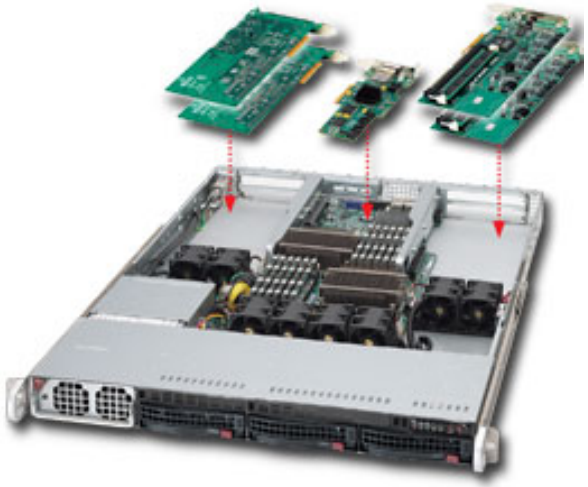
Lawrence Livermore (LLNL)

- Hyperion data intensive test-bed
- US Nuclear Security Administration
  - Advanced simulation and computing program
- >1100 nodes
  - ~100 teraflop compute capacity
- Over 9TB memory
- Over 100TB flash memory

LLNL Deployment & Results

Supermicro Servers

- 80 servers, 2 42U racks



Specs

- Processor 5600/5500
- Up to 192GB DDR3
- 4x PCI-E 2.0 (x8) slots
- 2 MLC ioDrive Duos (320GB per Duo)
- 2 ioSANs (320GB per ioSAN)

Performance

With 160 initiators:

40,800,000 sustained IOPS, 400GB/sec sustained bandwidth

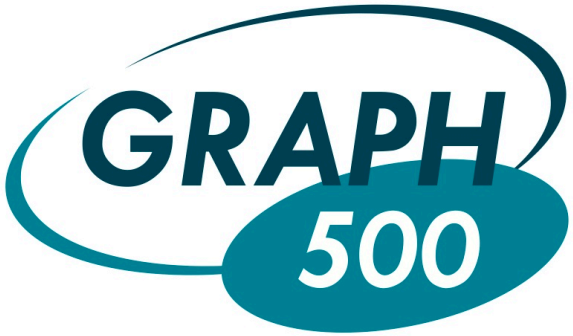
Transparent Memory Tiering

- Transparent memory tiering enables application use of flash via memory semantics, without application modifications
- Alternative to storage based tiering (transparent or otherwise)
- Limited by existing OS swap implementations
  - Performance, concurrency
- TEAM – Transparent Extension of Application Memory – user level paging and flash aware optimization *
  - Result >=4x SWAP performance

* Under conference submission

Graph500

- Traversing massive graphs
  - "Using 2.56TB of Fusion-io NAND flash to access data using memory semantics, LLNL's new Graph500 algorithm can process graphs 8x larger than before with only a 50% performance degradation compared to an all DRAM system."
- Results: 55.6 MTEPS* [2]
  - 4 x 640GB Fusion-io MLC



* Million Traversed Edges per Second

[2] Graph500: Traversing massive graphs with NAND flash; Pearce, Gokhale, & Amato

Critical Reliability Elements

- Validation in a variety of server platforms
- Simplified data path
- Sophisticated error correction
- Ability to handle flash errors or flash failures seamlessly
- Field upgradable products
- Monitoring and management

Is Disk Dead?

- No, but
  - A new hierarchy – performance storage needs met by flash (and futures) as new tier
    Significant disk and DRAM displacement
  - Server side flash
  - Disk moves to capacity focus, massive data, archival, power managed

References

- DFS – A File System for Virtualized Flash Storage, FAST 2010
- Beyond Block I/O – Rethinking Traditional Storage Primitives, HPCA 2011
- SSDAlloc – Hybrid SSD/RAM Management Made Easy, NSDI 2011
- Multithreaded Asynchronous Graph Traversal for In-memory and Semi-Extended-Memory, SC 2010
- PTRIM + EXISTS – Exposing New FTL primitives to Applications, UCSD NVM Workshop 2011
- TEAM – Transparent Extension of Application Memory – Under submission

ntalagala@fusionio.com